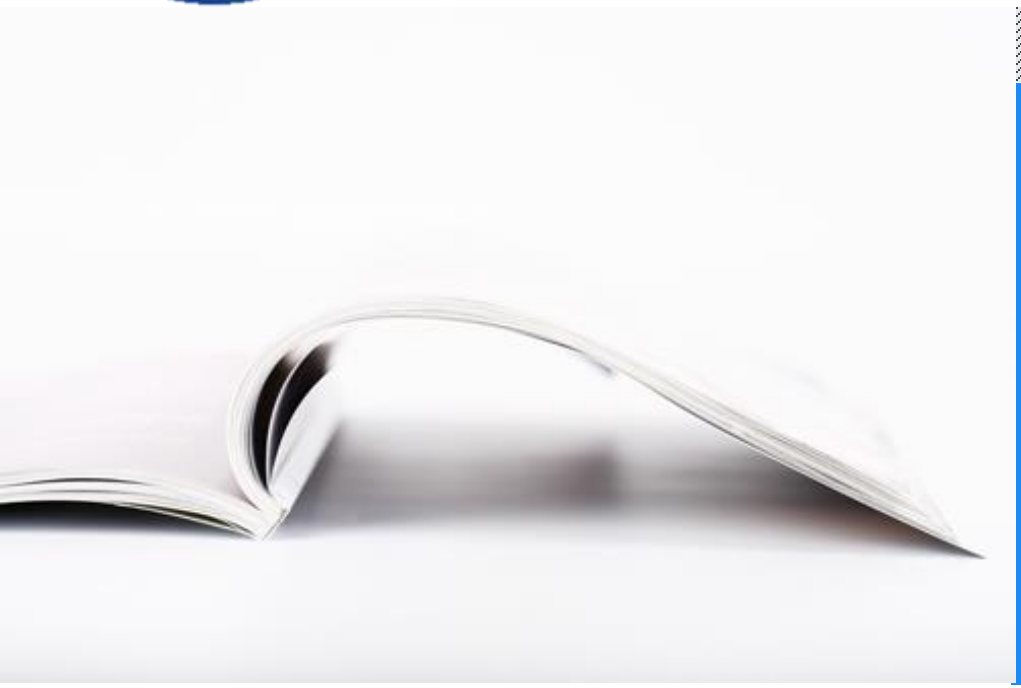


Web信息处理与应用

第十二节

知识抽取与表达（上）

徐童

2023.11.20

上节回顾

- **信息抽取的内容**

- 核心的8字方针：“**抽取实体，确定关系**”
  - **实体**：即命名实体，指文本中的基本构成块，如人、机构等
  - 属性：实体的特征，如人的年龄、机构的类型等
  - 关系：实体之间存在的联系，也称事实，如公司和地址之间的位置关系、公司与人之间的雇佣关系
  - 事件：实体的行为或实体参与的活动

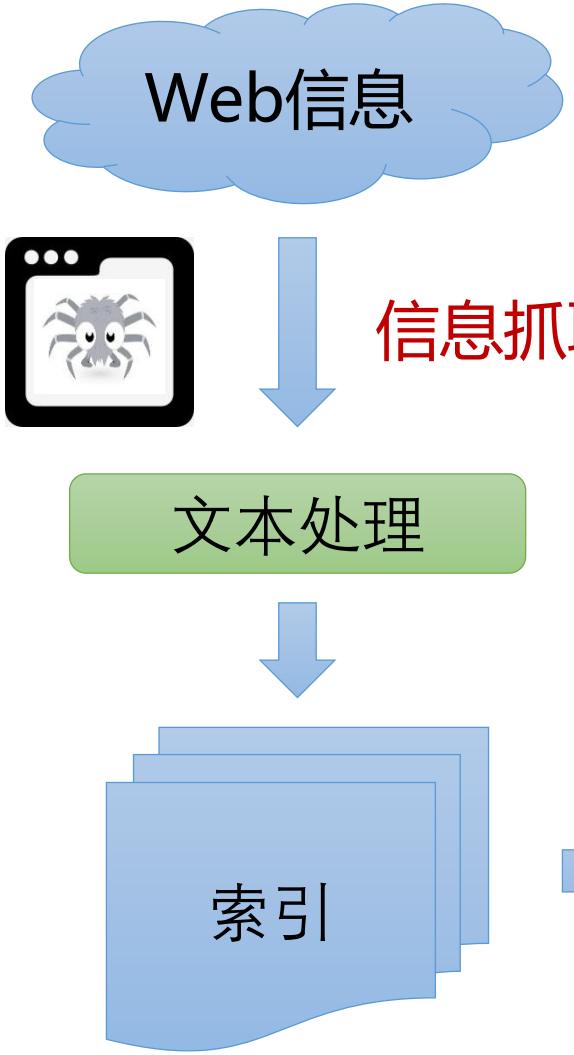
上节回顾

- **信息抽取的基本任务**

- **命名实体NE**（实体抽取）
- 命名实体抽取是信息抽取最重要的任务
- 命名实体是文本中基本的信息元素，是正确理解文本的基础
  - 狭义：指现实世界中具体或抽象的实体，如人、组织、地点等
    - 如：水果摊/Org，老板 郝哥/Person
  - 广义：还可以包含日期和时间、数量表达式等

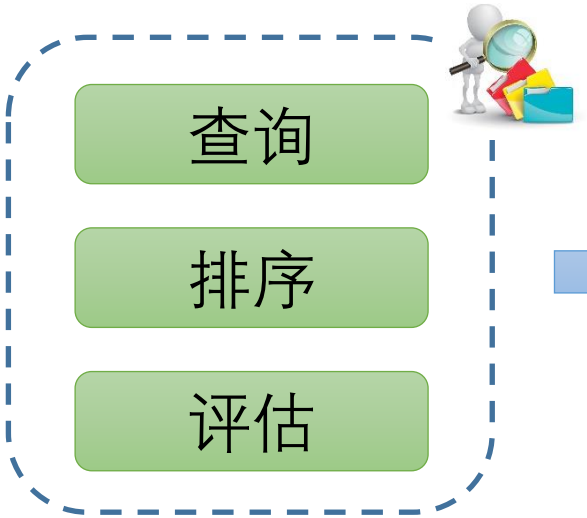
本节主题

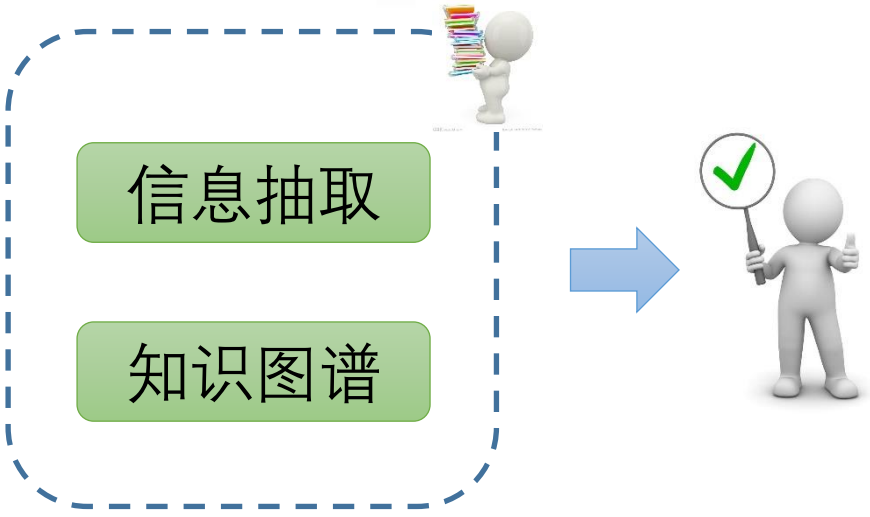
- 本课程所要解决的问题



第十一个问题：

如何有效抽取和表达实体？





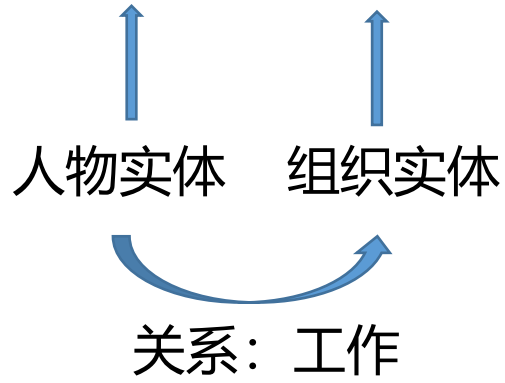
本节目录

- **实体抽取任务**
  - 任务定义
  - 基本方法
- 实体对齐
- 实体链接

- **命名实体识别的基本概念**

- 命名实体识别（Named Entity Recognition，NER）
  - 识别出文本中的人名、地名等专有名称，和有意义的时间、日期等数量短语等，并加以归类。
  - 命名实体识别是信息抽取中的核心任务，它往往包含两个子任务
    - 判别实体**边界**
    - 判别实体**类型**

- 例如：比尔盖茨在微软工作。



- **命名实体识别的内容**

- 一般按照MUC-7的定义，分为3大类，7小类
  - 实体类：人名、地名、机构名
  - 时间类：日期、时间
  - 数值类：货币、百分比
- 在MUC-7的严格定义下，哪些不属于命名实体？（部分例子）
  - 重复指代的普通名词：如 飞机、公司 等
  - 人的团体名称以及以人命名的法律、奖项等：如 共和国、诺贝尔奖 等
  - 非时间、日期、货币、百分比的数字

ACE (Automatic Content Extraction)定义中的NER任务：
人名（Person）、机构名（Organization）、地名（Location）、设备名（Facility）、武器名（Weapon）、交通工具名（Vehicle）和地理政治实体（Geo-Political Entity）

- **较为特殊的实体识别任务**

- 数值识别与检测任务
  - 百分比：[25/100]
  - 钱：[20 元]
  - 邮箱： *example@domain.com*
  - 时间：美国商会中国分会[近日]派出一个25人组成的代表团，在华盛顿向国会和白宫展开为期[一周]的游说活动

- 时间表达识别与检测任务
  - [1995年10月]，中国长江三峡工程开发...
  - [公元前2100年左右]，美索布达米亚人已有了乘法表。

- **命名实体识别的难点**

- 与分词的难点非常相似
  - 不断有新的命名实体涌现，如新的人名、地名、组织名等
  - 命名实体存在严重歧义
    - 如Washington（地名/人名），May（人名/月份）
  - 命名实体构成结构复杂，如别名、缩略词、音译、包含数值的实体等
    - 如USTC 与 Univ. Sci. Tech. of China，1958年9月20日 等
  - 命名实体类型多样：如John Smith，Mr Smith, John，实际上是共指关系

- **命名实体识别的性能评价**

- 与检索任务大致相同，采用Precision / Recall / F-value加以衡量
- 正确率与召回率的计算方式
  - 方案1：分子为返回的正确答案数量
  - 方案2：分子为返回的正确答案数量 + ½的部分正确答案数量
    - 部分正确的案例：“Severus/Person Snape”（类型正确，边界错误）

- **命名实体识别方法（1）：基于词典**

- 基于词典的识别方法（List Lookup）
  - 经常作为NER问题的基准算法（baseline）
  - 预先构建一个命名实体词典，出现在词典中的词汇即识别为命名实体
  - 词典的来源：来自于领域公开数据，例如
    - 人名/组织名，可以来自于黄页、电话簿、公开名单等
    - 地点，可利用一些现有的地理信息列表

- **命名实体识别方法（1）：基于词典**

- 基于词典的识别方法（List Lookup）
  - 优点：简单快速，与具体语境/领域无关，容易部署和更新（只需更新词典）
  - 缺点（与基于词典/匹配分词存在类似问题）：
    - 大部分情况下很难枚举所有的命名实体名
    - 构建和维护词典的代价较大
    - 难以有效处理实体歧义

• 命名实体识别方法（2）：基于规则

- 采用手工构造规则模板，对符合规则的实体进行识别
  - 选用特征包括统计信息、标点符号、关键字、指示词和方向词、位置词（如尾字）、中心词等
  - 以模式和字符串相匹配为主要手段
    - 例如：报告人：[漆桂林] 教授 （[东南大学]）

人物实体　　组织实体

    - 可借助结构/半结构化或固定表达生成模板

- **命名实体识别方法（2）：基于规则**

- 采用手工构造规则模板，对符合规则的实体进行识别
  - 基于规则模板的方法在深度学习技术发展之前被广泛使用，多数参加MUC-7 （1997）会议评测的系统，都采用了此方法
  - 例如：[组织] 位于 [位置] 的总部
    - 我们访问了[华为]位于[东莞]的总部
  - 该类方法的局限性明显：不同句式意味着不同的模板
    - 例如，上述句式可改为：我们访问了华为的[总部]，它坐落于[东莞]

- **命名实体识别方法（2）：基于规则**

- 基于手工规则的识别方法
  - 优点：当提取的规则能较精确地反映语言现象时，性能较好
  - 缺点：
    - 不同表达对应不同规则，导致规则库极其庞大，使用不便
    - 规则往往依赖于具体语言、领域和文本风格

      不同领域的句法往往差异极大，如学术圈与二次元
    - 代价太大，系统建设周期长、移植性差而且需要建立不同领域知识库

- **命名实体识别方法（3）：基于统计**

- 基于统计的命名实体识别方法是当下的主流方法

类型	采用的模型或方法	代表工作
有监督的学习方法	隐马尔可夫模型或语言模型	Liu *et al.*(2005)；Zhang *et al.* (2003a)；Sun *et al.* (2002)；Zhou and Su(2002)；Bikel *et al.*(1997)
	最大熵模型	Tsai *et al.*(2004)；Borthwick(1999)；Mikheev *et al.*(1998)
	支持向量机	Yi *et al.*(2004)；Asahara and Matsumoto (2003)
	条件随机场	Leaman and Gonzalez (2008)；Finkel *et al.* (2005)；McCallum and Li (2003)
	决策树	Isozaki(2001)；Paliouras *et al.*(2000)；Sekine *et al.*(1998)
半监督的学习方法（弱监督学习方法）	利用标注的小数据集（种子数据）自举学习	Singh *et al.*(2010)；Nadeau(2007)；Niu *et al.* (2003)；Collins (2002b)；Collins and Singer (1999)
无监督的学习方法	利用词汇资源（如 WordNet）等进行上下文聚类	Etzioni *et al.*(2005)；Shinyama and Sekine (2004)
混合方法	几种模型相结合或利用统计方法和人工总结的知识库	Liu *et al.*(2011b)；Finkel and Manning(2009)；Zhou(2006)；Wu *et al.*(2003，2005)；Jansche and Abney(2002)

先前回顾

- **回顾：分词时的序列标注问题**

- 基于统计模型的分词方法，进一步抽象而言，可以得到一个序列标注问题
  - 四类标注：B（词的开始）、M（词的中间）、E（词的结束）、S（单字词）
  - 例子：中国科学技术大学是中国最好的大学
    - 标注：BMMMMMME S BE BME BE
    - 分词结果：中国科学技术大学 / 是 / 中国 / 最好的 / 大学

- 类似的序列标注，在命名实体识别问题中也得到广泛应用。

- **命名实体识别方法（3）：基于统计**

- 分支一：基于分类的命名实体识别方法
- 将NER视作一个多分类问题，通过设计特征训练分类器的方法加以解决。
- 例如：Hideki Isozaki, et al., Efficient Support Vector Classifiers for Named Entity Recognition, COLING 2002
  - 一共33个标签：8种实体，每种对应Begin, Middle, End, Single四种类型，加上Other（即不属于任何一类实体），得到8x4+1 = 33类标签。
  - 选取15维特征：当前词及前后各两个词（共计5个词）
    - 每个词3维特征：词性、字符类型、单词

- **命名实体识别方法（3）：基于统计**

- 通过以上方式，得到一个高维稀疏的向量，仅15维为1，其余均为0
- 特征实例：对于“President George Herbert Bush said Clinton is ...”中“Bush”这个词

$$
\begin{aligned}
x[1] &= 0 \quad // \text{ Current word is not 'Alice'} \\
x[2] &= 1 \quad // \text{ Current word is 'Bush'} \\
x[3] &= 0 \quad // \text{ Current word is not 'Charlie'} \\
&\vdots \\
x[15029] &= 1 \quad // \text{ Current POS is a proper noun} \\
x[15030] &= 0 \quad // \text{ Current POS is not a verb} \\
&\vdots \\
x[39181] &= 0 \quad // \text{ Previous word is not 'Henry'} \\
x[39182] &= 1 \quad // \text{ Previous word is 'Herbert'} \\
&\cdot
\end{aligned}
$$

- 基于该向量，通过支持向量机（SVM）模型+Sigmoid函数属于何种标签

- **命名实体识别方法（3）：基于统计**

- 事实上，早期基于统计的方法，需要精心设计大量的相关特征
  - 以机构名识别为例，常见的内部特征包括单词特征、核心词特征、词性特征、语义特征等

标注	类型	示例
F	机构特征词	北京搜狐畅游时代网络技术***有限公司***
R	机构名中的人名	法国***马蒂尼埃***集团
NR	其它人名	***俞昊然***创立了“计蒜客”
S	机构名中的地名	***北京市***文化局相关领导表示
NS	其它地名	在前不久的***中国***游戏行业年会上
O	常见机构名	***中国人民银行***
E	机构名中的其它词	侵犯***腾讯***公司相关游戏著作权一案
L	机构名之间的连接词	中国移动***和***中国联通慢慢掌控了很多版权
P	职位名	友达***董事长***李焜耀
Z	其它词	
……	……	……

- **命名实体识别方法（3）：基于统计**

- 与统计特征相关的一个问题：**词性标注问题**
  - 词性（part-of-speech）是词汇基本的语法属性，通常也称为词类。
  - 词性标注就是在给定句子中判定每个词的语法范畴，确定其词性并加以标注的过程。词性标注是自然语言处理中一项非常重要的基础性工作。
- 词性标注问题，尤其是中文词性标注问题，也面临一些困难和挑战：
  - 汉语是一种缺乏词形态变化的语言，无法从单词形态上来判别
  - 常用词兼类现象严重，例如：科学技术（名词）/ 这不科学（形容词）

- **命名实体识别方法（3）：基于统计**

- 基本的**词性标注方法**
  - 从思路上说，词性标注方法与实体识别，乃至分词，总体思路都是类似的
  - 基于规则的方法：人工或通过大规模语料学习规则
    - 核心思想是按兼类词搭配关系和上下文语境建造词类消歧规则
  - 基于统计模型的标注方法：HMM等面向词序列的方法
  - 统计方法与规则方法相结合的词性标注方法
    - 先基于规则排除明显歧义，再基于统计模型标注，最后人工校验

➢ 可参考统计自然语言处理（第7.5节），宗成庆著，北京大学出版社

- **命名实体识别方法（3）：基于统计**

- 顺带一提：B、M、E、S、O五类标签可能过多，尤其是在实体种类较多时
- 目前，更为简化的是BIO标签体系,，其中
  - I 合并了M、E所承担的功能，即仅区分开头和后缀，不区分中间和结尾
  - 单字实体（S）用没有任何 I 后缀的 B 取代
- 实例：

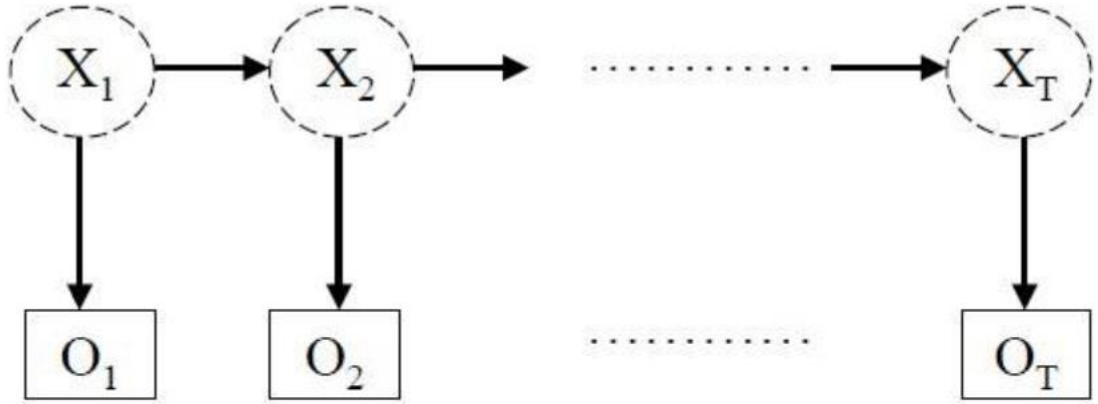
中	国	科	学	技	术	大	学	的	徐	童	老	师	在	上	课
B-ORG	I-ORG	I-ORG	I-ORG	I-ORG	I-ORG	I-ORG	I-ORG	O	B-PER	I-PER	O	O	O	O	O

- **命名实体识别方法（3）：基于统计**

- 基于统计的命名实体识别方法具有以下特点：
  - 对特征选取的要求较高，需要从文本中选择对NER有影响的特征来构建特征向量（尤其是早期工作，深度学习技术发展后相对要求降低）
  - 通常做法是对训练语料所包含的语言信息进行统计和分析，从中挖掘出特征
  - 对语料的依赖也较大，目前缺少通用的大规模语料
    - 对深度学习技术影响尤甚，特定专业领域影响最为明显
  - 大部分技术仍需要进行人工标注训练数据

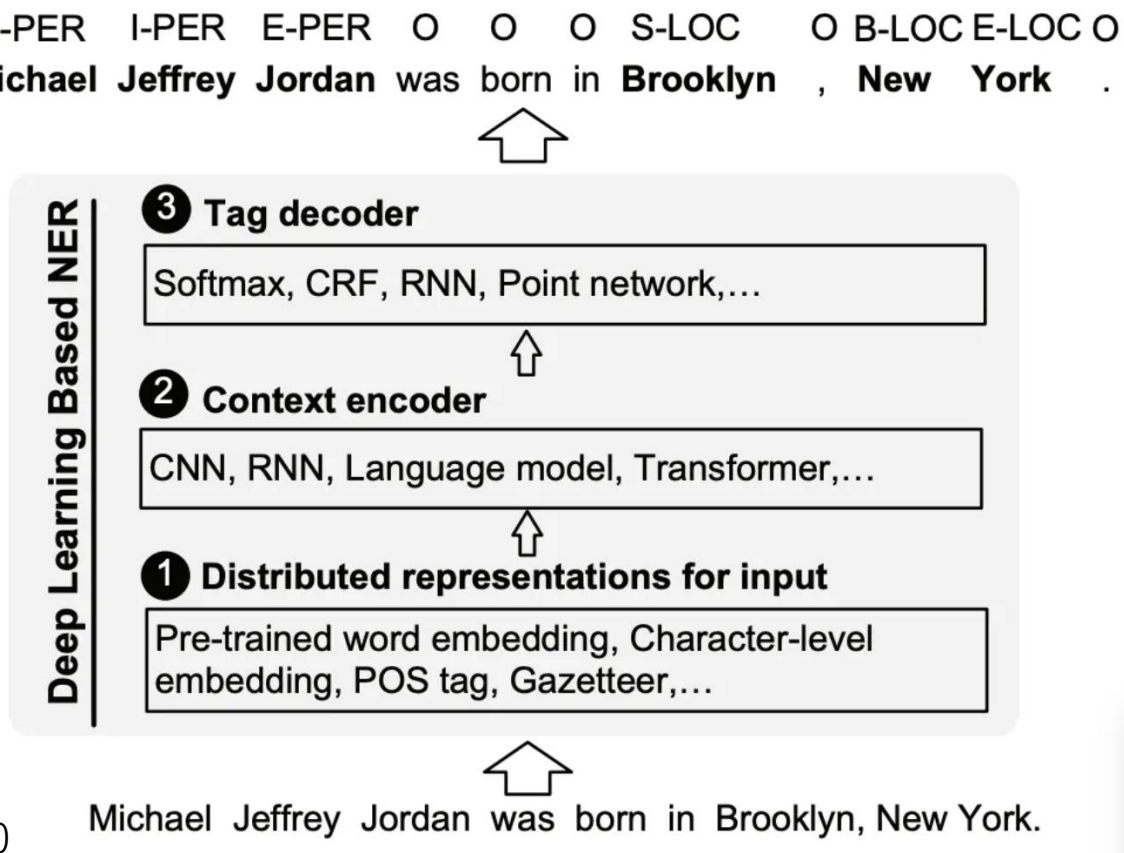
- **命名实体识别方法（3）：基于统计**

- 分支二：基于序列模型的命名实体识别方法
- 与分词中的序列标注方法思路类似，区别在于标注的不同
  - 针对命名实体的类别不同，引入了更多、更细致的标签种类
  - 常用模型亦采用HMM、CRF以及各种序列深度学习方法（如LSTM）等



- **命名实体识别方法（3）：基于统计**

- 分支二：基于序列模型的命名实体识别方法
- 深度学习方法的一般框架如下图所示：
  - 第一步，将单词和对应的标签实现初步（分布式）向量化
  - 第二步，基于编码器对文本进行表征，其中通过RNN等方法捕捉上下文
  - 第三步，基于解码器进行分类



A Survey on Deep Learning for Named Entity Recognition, TKDE 2020

- **命名实体识别方法（3）：基于统计**

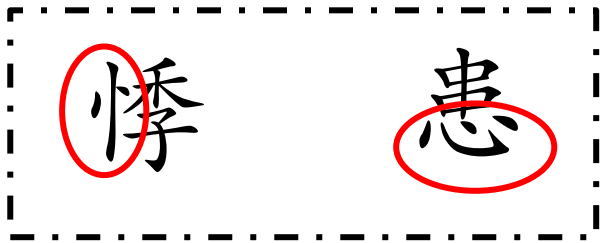
- 除了模型的改进之外，研究者们也在尝试引入更多信息和领域知识以提升效果
  - 例如，英文的词根词缀信息，或者汉字的部首信息，都蕴含着丰富语义
  - 一个医疗实体识别的案例：中文五行等特征部首常出现在医疗实体中
    - 不仅是实体开始的提示符，不同类型的部首，还往往对应着不同类型的医疗实体

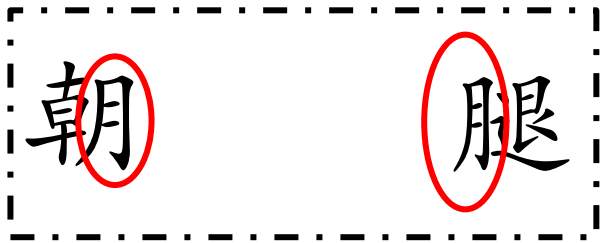
- 金：人体内微量元素
- 木：中成药物
- 水：人体体液与体液症状（溶、溃…）
- ……

- 月：身体部位（腿、胃、肾…）
- 口：头部器官
- 疒：疾病名称
- ……

- **命名实体识别方法（3）：基于统计**

- 如何利用部首信息进行建模？三点观察
  - 简体中文汉字往往都是由繁体中文汉字简化演变而来，因而简体部首可能不具有很强的解释性。
  - 存在不同部首表达同一种释义的情况：“阝”与“耳”、“忄”与“心”、“火”与“灬”等。增大了训练代价，也降低了部首信息的表达能力。
  - 释义不同的字符有着相同的部首：“朝”(释义为早晨)与“脚(释义为足)”的部首都是“月”。同样影响了部首的表达能力。





- **命名实体识别方法（3）：基于统计**

- 基于前述观察，通过采用繁体部首的方式，还原语义信息

『悸』
拼音：jì　注音：ㄐㄧˋ
简体部首：忄　部首笔画：3　总笔画：11
繁体部首：心　部首笔画：4　总笔画：12
康熙字典笔画(悸:12；)
五笔86：NTBG　五笔98：NTBG　仓颉：PHDD
四角号码：92047　UniCode：U+60B8　规范汉字编号：4988

『腿』　异体字：骽踠
拼音：tuǐ　注音：ㄊㄨㄟˇ
简体部首：月　部首笔画：4　总笔画：13
繁体部首：肉　部首笔画：6　总笔画：16
康熙字典笔画(腿:16；)
五笔86：EVEP　五笔98：EVPY　仓颉：BYAV
四角号码：77233　UniCode：U+817F　规范汉字编号：2976

释义相同，在繁体部首上统一

释义不同，在繁体部首上区分开

『患』
拼音：huàn　注音：ㄏㄨㄢˋ
部首：心　部首笔画：4　总笔画：11
康熙字典笔画(患:11；)
五笔86：KKHN　五笔98：KKHN　仓颉：LLP
四角号码：50336　UniCode：U+60A3　规范汉字编号：2285

『朝』
拼音：cháo　zhāo　注音：ㄔㄠˊ　ㄓㄠ
部首：月　部首笔画：4　总笔画：12
康熙字典笔画(朝:12；)
五笔86：FJEG　五笔98：FJEG　仓颉：JJB
四角号码：47420　UniCode：U+671D　规范汉字编号：2559

• 命名实体识别方法（3）：基于统计

- 相应的，基于部首信息，对模型进行改进
- 基础模型：采用LSTM+CRF的方式实现医疗命名实体识别
  - 单纯LSTM忽略了标签序列的关联性，CRF将提升标签序列的合理性
- 部首信息对于模型的改进体现在以下两个方面
  - LSTM部分，在字向量编码中加入部首编码，与字符向量拼接来表示字符
  - CRF部分，加入部首标签矩阵，区分不同部首对应不同类型的不同可能性

➢ 李丹等，部首感知的中文医疗命名实体识别，中文信息学报，2020

- **命名实体识别方法（3）：基于统计**

- 实验证实，引入部首信息之后，在医疗命名实体识别任务上取得了更好效果
- 案例1：“患者肺部轻度慢性发炎”
  - 实体嵌套，传统方法容易拆成多个实体对待，导致实体支零破碎
  - 基于部首信息，可抽取出以“肺”为开始，“炎”为结尾的实体词
- 案例2：“术后予头孢美锉钠抗感染，止血，及补液等对症支持治疗。”
  - 传统方法会将“头孢美锉钠抗感染，止血，及补液”作为一个完整的实体
  - 基于部首信息，可以根据“钠”的部首判断其作为一个药品实体词的结尾

- **命名实体识别的常用工具**

- 英文命名实体识别的常用工具
  - Stanford NER
    - 斯坦福大学开发的基于CRF的NER系统，基于CoNLL、MUC-7和ACE等语料训练
    - https://nlp.stanford.edu/software/CRF-NER.shtml
  - MALLET
    - 麻省大学开发的统计自然语言处理的开源包，其序列标注工具的应用中能够实现NER。
    - http://mallet.cs.umass.edu/

- **命名实体识别的常用工具**

- 中文命名实体识别的常用工具
  - NLPIR-ICTCLAS：https://github.com/NLPIR-team/NLPIR

    介绍中文分词时提到的可视化分词工具，同时可实现词性判别与实体识别
  - HanLP：https://github.com/hankcs/HanLP

    一系列模型与算法组成的NLP工具包，支持命名实体识别
  - NLTK：http://www.nltk.org/

    一个高效的Python构建的平台，用来处理人类自然语言数据。

- **与命名实体识别相关的问题：实体消歧**

- 实体消歧（Entity Disambiguation），本质在于一个单词很可能有多个意思
- 这就意味着，在不同的上下文中所表达的含义可能不太一样。
  - 例如，介绍查询意图的歧义问题时提及的“苹果”

id	实体名	实体描述
1001	苹果	美国一家高科技公司，经典的产品有Iphone手机
1002	苹果	水果的一种，一般产自于…
…	…	…

- **与命名实体识别相关的问题：实体消歧**

- 解决实体歧义问题，首先需要获取实体的各种不同含义
  - 首先，对不同的含义抽取其相关内容，如描述文本，并建立关键词表
  - 其次，通过对关键词表的语义分析，从中抽取和归并相应的“概念”
    - 例如，苹果（水果）可能对应“富士”、“烟台”等，而苹果（手机）可能对应“iPhone”、“刘海屏”等。
  - 最终，对关键词进行语义表征，得到不同语义的表征向量。
- 由此，可以通过语义相似性（如余弦相似度）判断究竟属于哪种语义的实体。

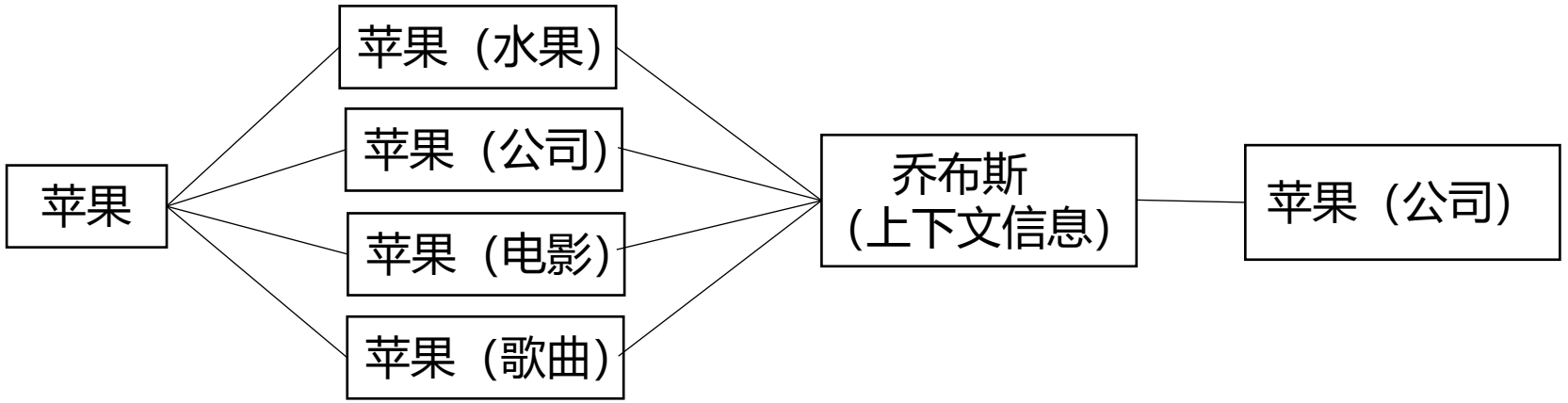
本节目录

- 实体抽取任务
  - 任务定义
  - 基本方法
- **实体对齐**
- 实体链接

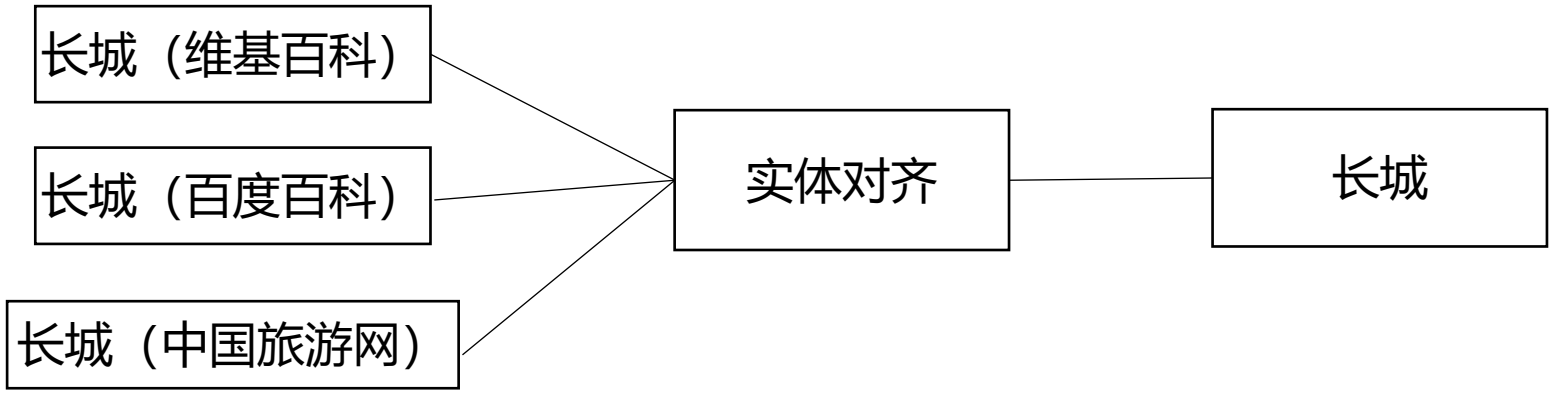
- **实体对齐**

- 实体对齐（Entity Alignment），也称实体匹配（Entity Matching）
- 指对于异构数据源知识库中的各个实体，找出属于现实世界中的同一实体。
  - 例如，不同药物可能在不同数据库中采用不同的名称
    - E.g., 利君沙（琥乙红霉素片）
- 一般而言，利用实体的属性信息判定不同源实体是否可对齐

- **实体对齐与实体消歧的区别**
- 实体消歧旨在消除一词多义的歧义现象

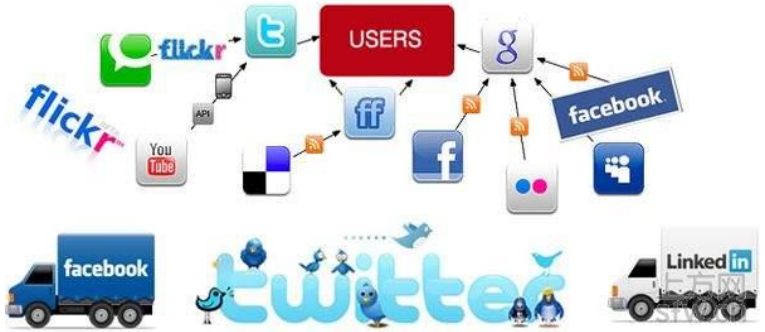


- 实体对齐旨在表征同一对象的多个实体之间构建对齐关系，丰富实体信息



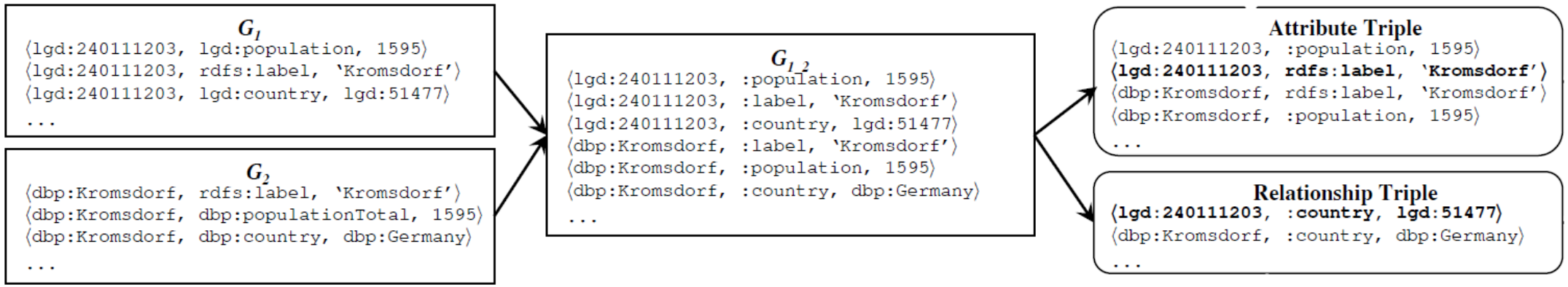
- **基础任务：基于表征的知识图谱实体对齐**

- 近来，针对跨知识图谱（KGs）的实体对齐任务，研究者提出并改进了多种基于表征（Embedding）的模型
  - 不仅利用实体的属性和语义信息，还利用实体间的关系
  - 要求KGs的表征（包括关系表征和实体表征）落在同一个向量空间
  - 换言之，这些模型更关注于关系三元组（relationship triple）
  - 一个类似的任务：跨社交网络用户匹配
    - 不仅考虑用户画像，也考虑社交关系相似性



- **基础任务：基于表征的知识图谱实体对齐**

- 如何使关系表征拥有统一的向量空间？利用相似性合并
  - 找到部分相似的谓词，例如，bornIn(KG1)与wasBornIn(KG2)，并用统一的命名方案（例如，bornIn）重新命名，将KG1和KG2合并为KG1_2
  - 合并后的图KG1_2被分割成关系三元组Tr和属性三元组Ta，用于表征学习



Bayu Distiawan Trisedya, et. al., Entity alignment between knowledge graphs using attribute embeddings, AAAI'2019, 297-304.

- **基础任务：基于表征的知识图谱实体对齐**

- 如何使实体表征拥有统一的向量空间？联合学习结构表征和属性字符表征
  - 在学习实体或关系表征时，参考了TransE的基本思想
    - TransE的核心假设：$\mathbf{h}+\mathbf{r}\approx\mathbf{t}$ （将在下节课再次详细介绍）
    - 为此，通过优化以下目标函数来训练各元素表征

$$J_{SE}=\sum_{t_r\in T_r}\sum_{t'_r\in T'_r}\max\left(0,[\gamma+f(t_r)-f(t'_r)]\right)$$

    - 其中 $f(t_r)=\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|$ ， $T_r{}'$ 为随机替换头/尾实体的伪三元组

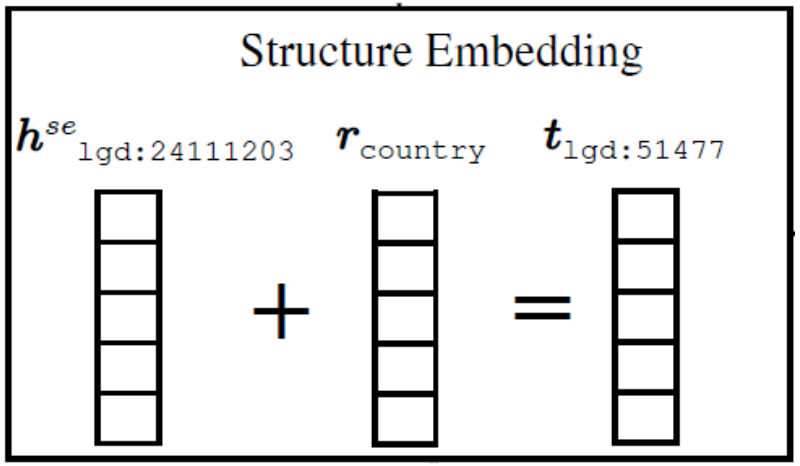
Bayu Distiawan Trisedya, et. al., Entity alignment between knowledge graphs using attribute embeddings, AAAI'2019, 297-304.

- **基础任务：基于表征的知识图谱实体对齐**

- 如何使实体表征拥有统一的向量空间？联合学习结构表征和属性字符表征
  - 使用TransE算法学习KG三元组的**结构表征**，增加权重α将表征学习的重点放在具有对齐谓词的三元组（通常对齐谓词出现数量更多）
  - 最小化以下目标函数：

$$J_{SE} = \sum_{t_r \in T_r} \sum_{t'_r \in T'_r} \max\left(0, \gamma + \alpha\left(f(t_r) - f(t'_r)\right)\right)$$

$$\alpha = \frac{count(r)}{|T|}$$

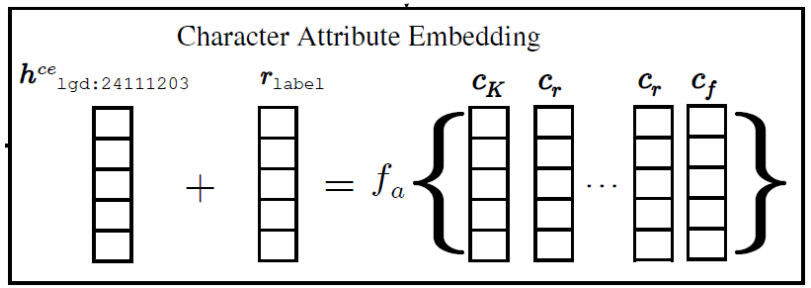


Bayu Distiawan Trisedya, et. al., Entity alignment between knowledge graphs using attribute embeddings, AAAI'2019, 297-304.

- **基础任务：基于表征的知识图谱实体对齐**

- 如何使实体表征拥有统一的向量空间？联合学习结构表征和属性字符表征
  - 沿用TransE的思路，将关系谓词r解释为从头部实体h到属性a的翻译
    - 使用组合函数$f_a(a)$来编码属性值
    - 由此，该部分的损失函数形式稍有区别：$f(t_a) = \|\mathbf{h} + \mathbf{r} - f_a(a)\|$
  - 为了学习**属性字符表征**，最小化以下目标函数：

$$J_{CE} = \sum_{t_a \in T_a} \sum_{t'_a \in T'_a} \max\left(0, [\gamma + \alpha\left(f(t_a) - f(t'_a)\right)]\right)$$



Bayu Distiawan Trisedya, et. al., Entity alignment between knowledge graphs using attribute embeddings, AAAI'2019, 297-304.

- **基础任务：基于表征的知识图谱实体对齐**

- 如何使实体表征拥有统一的向量空间？联合学习结构表征和属性字符表征
  - 沿用TransE的思路，将关系谓词r解释为从头部实体h到属性a的翻译
  - 几种可供选择的$f_a(a)$函数定义（其中的 c 均为属性中每个单词的表征）
    - 简单累加 $f_a(a) = \mathbf{c_1} + \mathbf{c_2} + \mathbf{c_3} + \ldots + \mathbf{c_t}$
    - 序列表征 $f_a(a) = f_{lstm}(\mathbf{c_1}, \mathbf{c_2}, \mathbf{c_3}, \ldots, \mathbf{c_t})$
    - N-gram $f_a(a) = \sum_{n=1}^{N} \left(\frac{\sum_{i=1}^{t} \sum_{j=i}^{n} \mathbf{c_j}}{t-i-1} \right)$

Bayu Distiawan Trisedya, et. al., Entity alignment between knowledge graphs using attribute embeddings, AAAI'2019, 297-304.

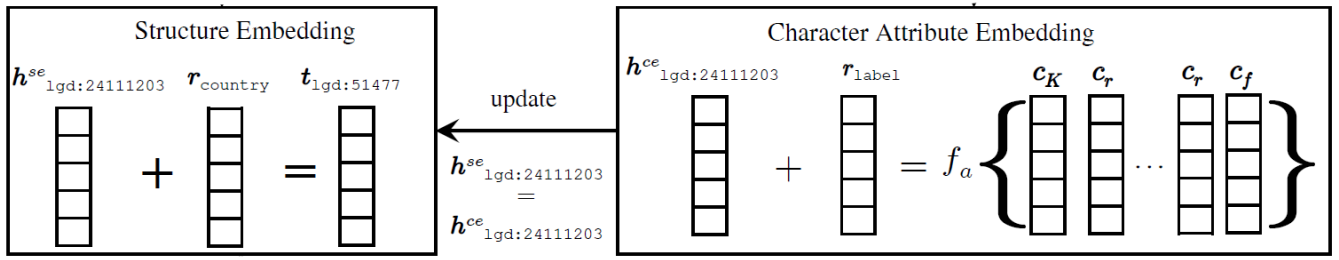
- **基础任务：基于表征的知识图谱实体对齐**

- 如何使实体表征拥有统一的向量空间？联合学习结构表征和属性字符表征
  - **结构表征**根据实体关系捕捉两个KG之间的实体的相似性，**属性字符表征**则是根据属性值来捕捉实体的相似性
  - 进一步实现属性字符表征h_{ce}与结构表征h_{se}的联合学习：

$$J_{SIM} = \sum_{h \in G_1 \cup G_2} [1 - \cos(\mathbf{h}_{se}, \mathbf{h}_{ce})]$$

  - 联合学习的总体目标函数是：

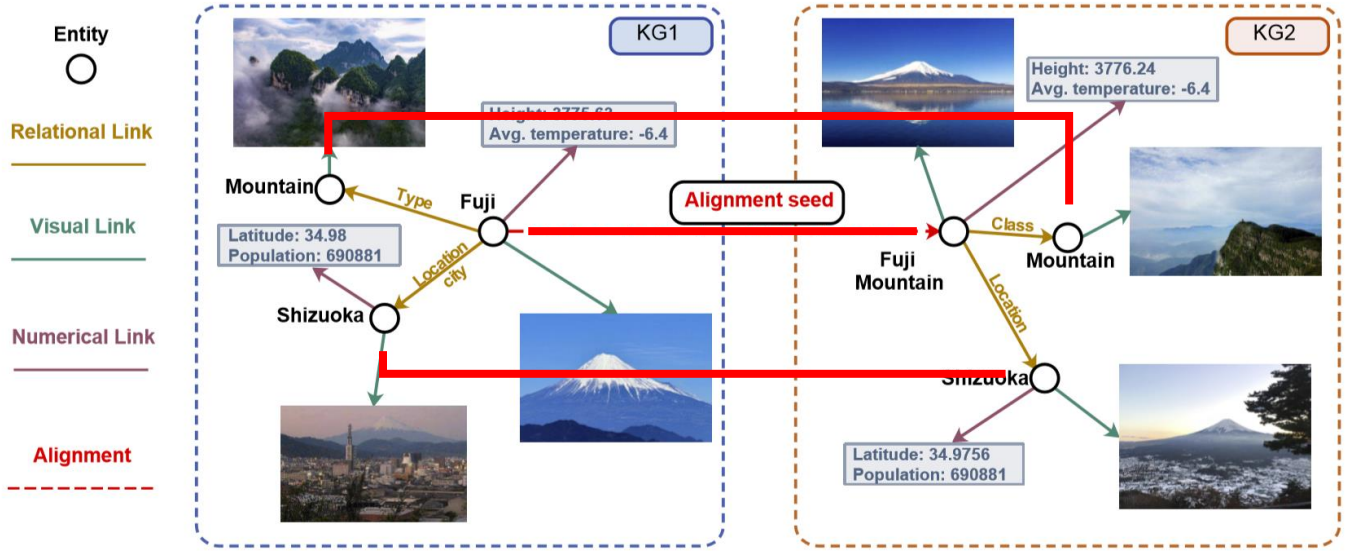
$$J = J_{SE} + J_{CE} + J_{SIM}$$



Bayu Distiawan Trisedya, et. al., Entity alignment between knowledge graphs using attribute embeddings, AAAI'2019, 297-304.

- **进阶任务：多模态知识图谱的实体对齐**

- 在真实场景中，知识通常被总结为各种形式，例如关系三元组、数字属性和图像等。使用不同的知识形式对图谱补全意义重大
  - 不仅需要利用实体的属性、语义和关系，还要有效处理其多模态特性

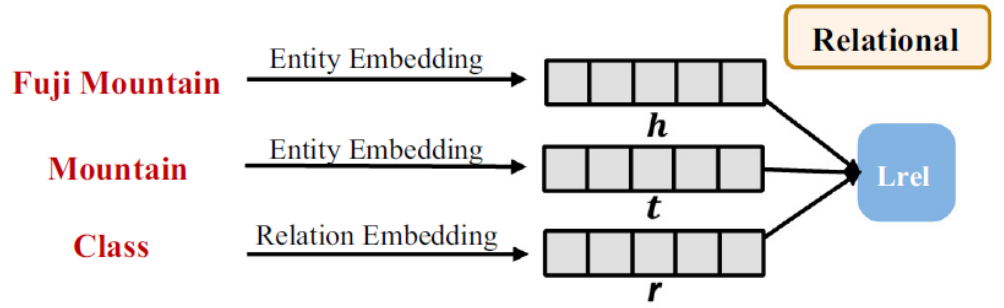


Liyi Chen, et. al., MMEA: Entity Alignment for Multi-Modal Knowledge Graphs, KSEM'20, 134-147.

- **进阶任务：多模态知识图谱的实体对齐**

- 如何表示不同类型的知识：通过多模态知识表征
  - **关系三元组**是KG的主要部分，对于判断不同KG的实体关联至关重要
  - 应用基于边界的损失函数，同时优化正负样本的得分
    - 参见以下公式，思路与先前AAAI 19工作类似，不再赘述

$$L_{rel} = \sum_{\tau^+ \in D^+} \sum_{\tau^- \in D^-} max(0, \gamma - f_{rel}(\tau^+) + f_{rel}(\tau^-)).$$



- 其中 $D^- = \left\{(h', r, t) \,|h' \in \widehat{E} \wedge h' \neq h \wedge (h, r, t) \in D^+ \wedge (h', r, t) \notin D^+\right\}$
  $\cup \left\{(h, r, t') \,|t' \in \widehat{E} \wedge t' \neq t \wedge (h, r, t) \in D^+ \wedge (h, r, t') \notin D^+\right\}.$

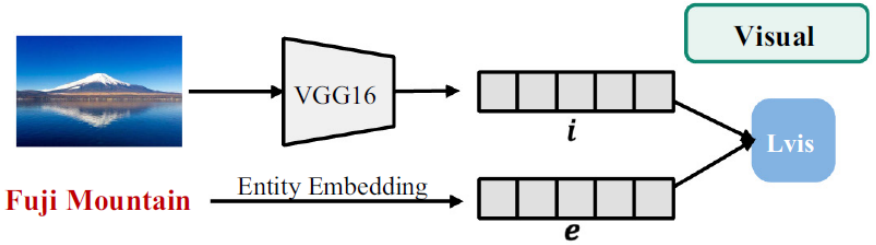
← 多了一个随机负采样不能采到正确样本的约束

Liyi Chen, et. al., MMEA: Entity Alignment for Multi-Modal Knowledge Graphs, KSEM'20, 134-147.

- **进阶任务：多模态知识图谱的实体对齐**

- 如何表示不同类型的知识：通过多模态知识表征
  - 视觉知识比关系知识更直观生动地反映了实体的外观，可以在一定程度上消除关系型信息的歧义
  - 对于**视觉类知识**，首先进行图像矢量化，并应用预先训练好的图像特征抽取模型，将图像矢量映射成对应的实体视觉向量
    - 用对应的实体表征限制实体视觉表征，并最小化以下损失函数：

$$L_{vis} = \sum_{(e^{(i)},i)\in Y} \log\left(1+\exp\left(-f_{vis}(e^{(i)},i)\right)\right)$$



Liyi Chen, et. al., MMEA: Entity Alignment for Multi-Modal Knowledge Graphs, KSEM'20, 134-147.

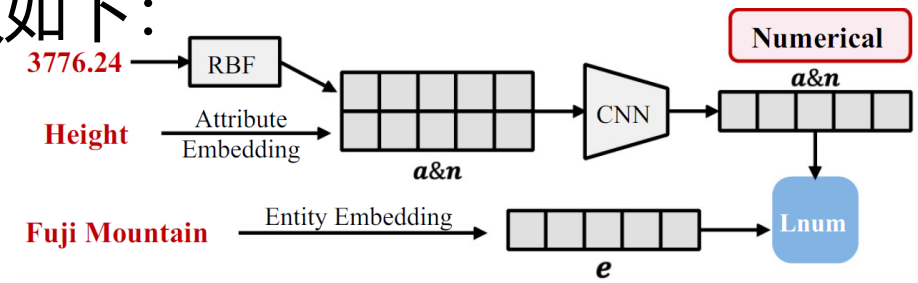
- **进阶任务：多模态知识图谱的实体对齐**

- 如何表示不同类型的知识：通过多模态知识表征
  - **数字属性**是无法构成关系事实的信息，可以作为关系知识的额外补充
  - 通过应用径向基函数将稀疏的数字信息转换为高维空间的嵌入：

$$\phi\left(n_{(e^{(n)},a_i)}\right) = \exp\left(\frac{-\left(n_{(e^{(n)},a_i)} - c_i\right)^2}{\sigma_i^2}\right)$$

  - 把属性的表征和数值向量连接起来，应用CNN网络提取其共同表征
  - 用对应的实体表征限制该共同表征，损失函数如下：

$$L_{num} = \sum_{(e^{(n)},a,n)\in Z} \log\left(1 + \exp\left(-f_{num}(e^{(n)}, a, v)\right)\right)$$

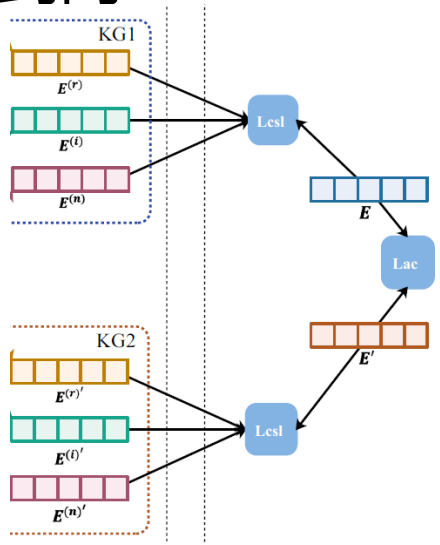


Liyi Chen, et. al., MMEA: Entity Alignment for Multi-Modal Knowledge Graphs, KSEM'20, 134-147.

- **进阶任务：多模态知识图谱的实体对齐**

- 如何利用多种模态表征完成实体对齐：通过多模态知识融合
- 来自不同模态的独立来源信息是往往是可以相互补充的，为更好的鲁棒性提供了额外的支持
  - 使用多模态知识融合模块来将异构的多种模态的知识表示映射到同一向量空间中

$$L_{csl}(\mathbf{E}, \mathbf{E}^{(r)}, \mathbf{E}^{(i)}, \mathbf{E}^{(n)}) = \alpha_1 ||\mathbf{E} - \mathbf{E}^{(r)}||_2^2 + \alpha_2 ||\mathbf{E} - \mathbf{E}^{(i)}||_2^2 + \alpha_3 ||\mathbf{E} - \mathbf{E}^{(n)}||_2^2$$



- 在训练过程中参数由L_{rel}，L_{vis}，L_{num}，L_{csl}共同更新

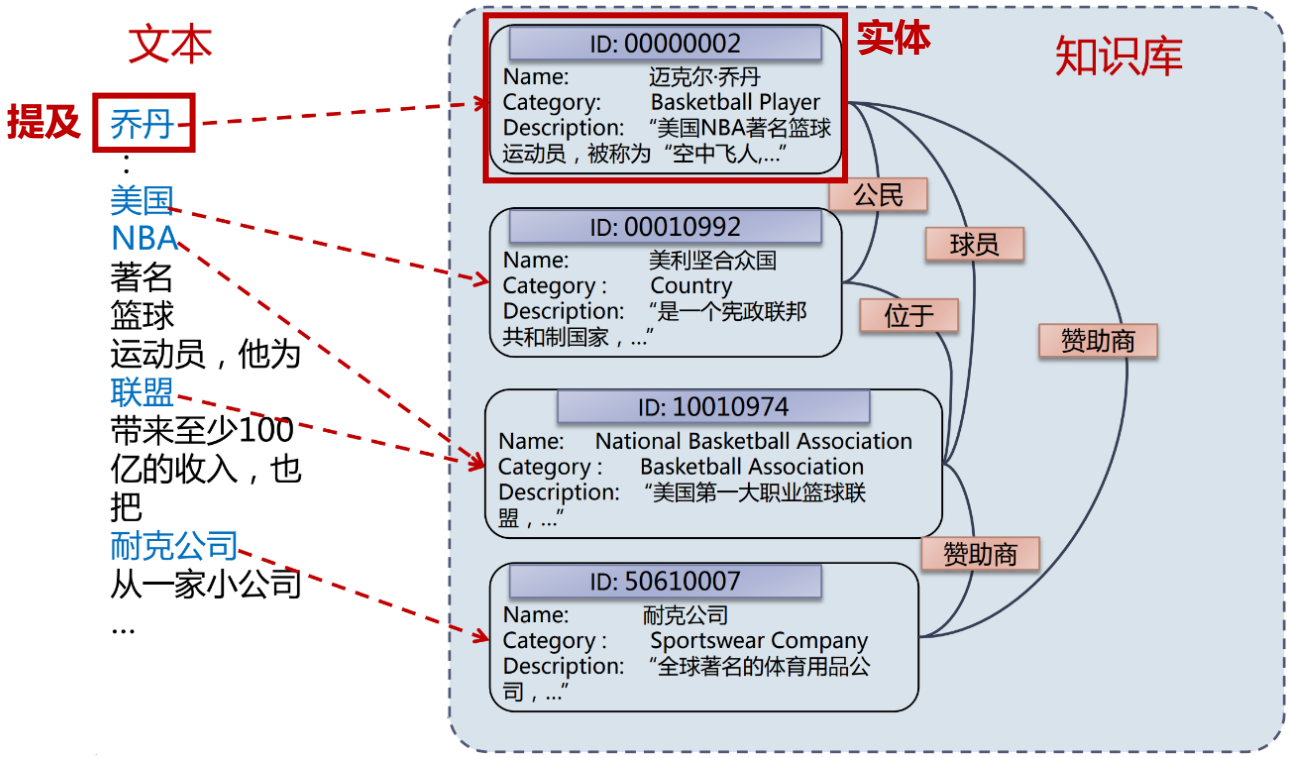
Liyi Chen, et. al., MMEA: Entity Alignment for Multi-Modal Knowledge Graphs, KSEM'20, 134-147.

本节目录

- 实体抽取任务
  - 任务定义
  - 基本方法
- 实体对齐
- **实体链接**

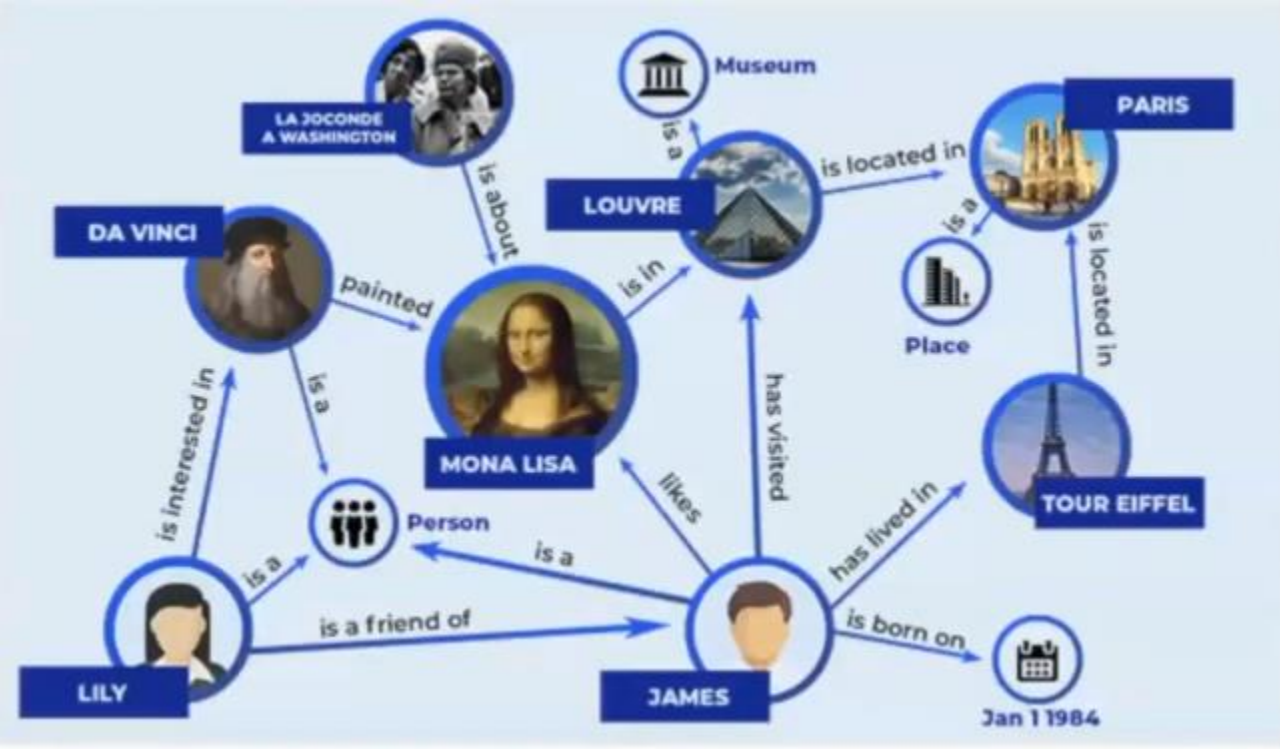
- **实体链接是什么?**

- 实体（entity），提及 or 指代（mention）
- 如何将文本中的提及链接到知识库中的实体上呢?



- **应用与挑战**

- 应用
  - 文本挖掘，知识图谱补全，信息检索和**问答系统**等

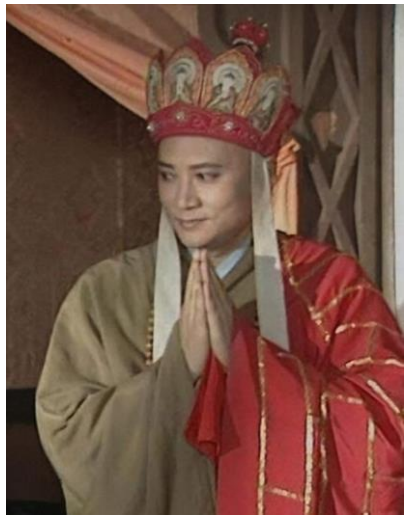


问答系统：实体链接是知识库问答（KBQA）的刚需，链接到实体之后才能查询图数据库。

蒙娜丽莎的作者是谁?
蒙娜丽莎在哪个博物馆?

- **应用与挑战**

- 挑战
  - 语言的**多样性**：同一个实体可能会有多个不同的提及
  - 语言的歧义性：同一个提及可能会对应多个不同的实体



唐僧
唐三藏
金蝉子
玄奘
长老

小说主要讲述了孙悟空出世，跟随菩提祖师学艺及大闹天宫后，遇见了**唐僧（唐三藏/玄奘）**、猪八戒、沙僧和白龙马，西行取经。

- **应用与挑战**

- 挑战
  - 语言的多样性：同一个实体可能会有多个不同的提及
  - 语言的**歧义性**：同一个提及可能会对应多个不同的实体

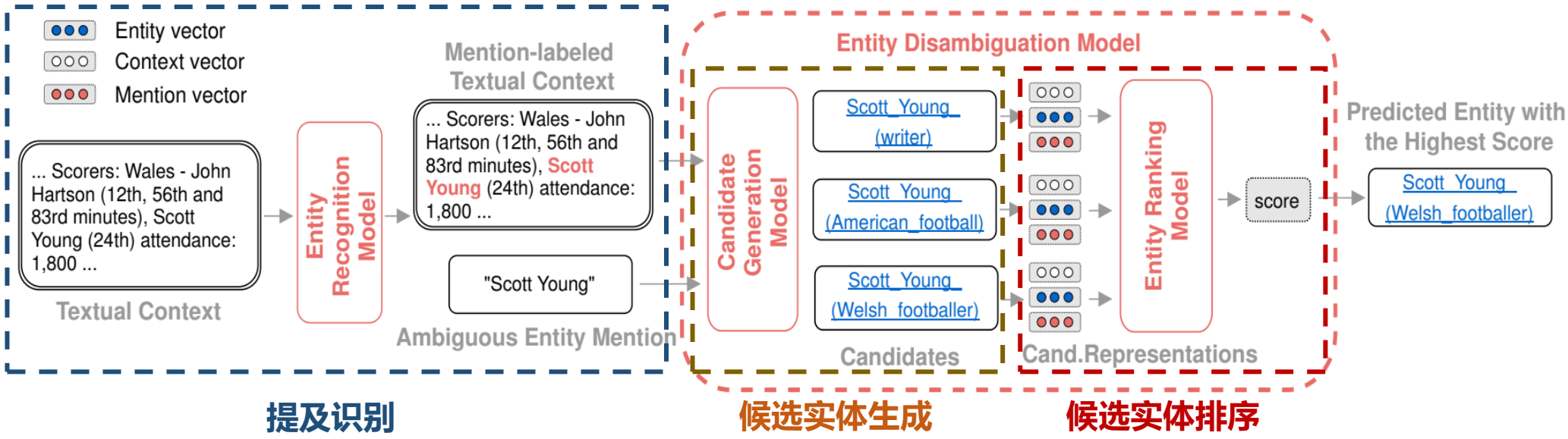
李娜 在澳洲公开赛战胜了齐布尔克娃

李娜，中国女子网球运动员，亚洲第一位大满贯女子单打冠军。

李娜，毕业于河南省戏曲学校，曾是中国大陆女歌手，出家后法名释昌圣。

- **任务框架**

- 提及识别：其本质与前述分词/实体识别类似
- 候选实体生成：粗粒度筛选，主要依赖字面匹配
- 候选实体排序：核心问题，主要基于语义分析



- **基于神经网络的实体链接技术**

- 代表性方法：Deep Joint Entity Disambiguation with Local Neural Attention (Ganea, et al. EMNLP 2017)
  - 对文档内包含的所有提及进行联合实体消歧
    - 局部实体消歧：一个提及和它的候选实体之间
    - 全局实体消歧：文档内的所有提及和候选实体组合之间

文档

提及1，（实体1，实体2，…，实体M）
提及2，（实体1，实体2，…，实体N）

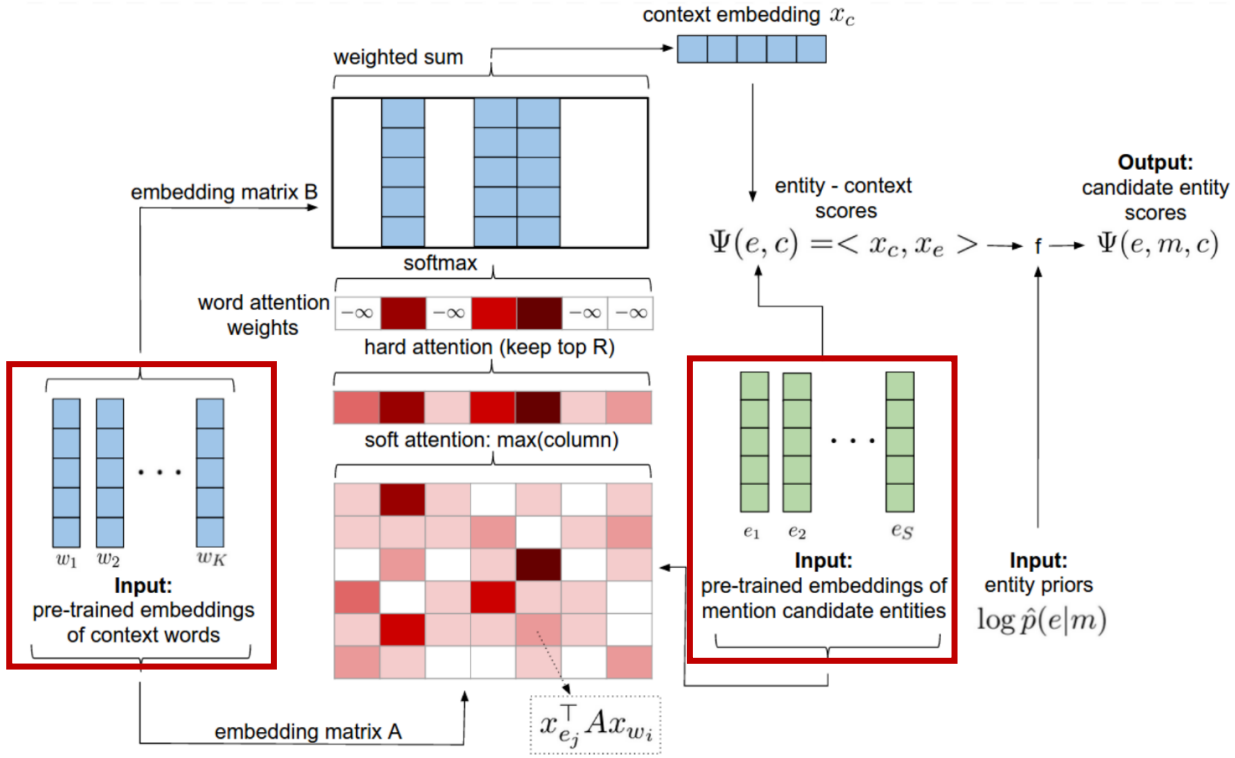
（提及1，提及2）
可能的候选实体组合：

（实体1，实体1）
（实体1，实体2）
…
（实体M，实体N）

- **基于神经网络的实体链接技术**

- 代表性方法：Deep Joint Entity Disambiguation with Local Neural Attention

  - 局部神经注意力计算局部分数



某个提及左右两侧K个单词组成的上下文：

$$c = \{w_1, \ldots, w_K\}$$

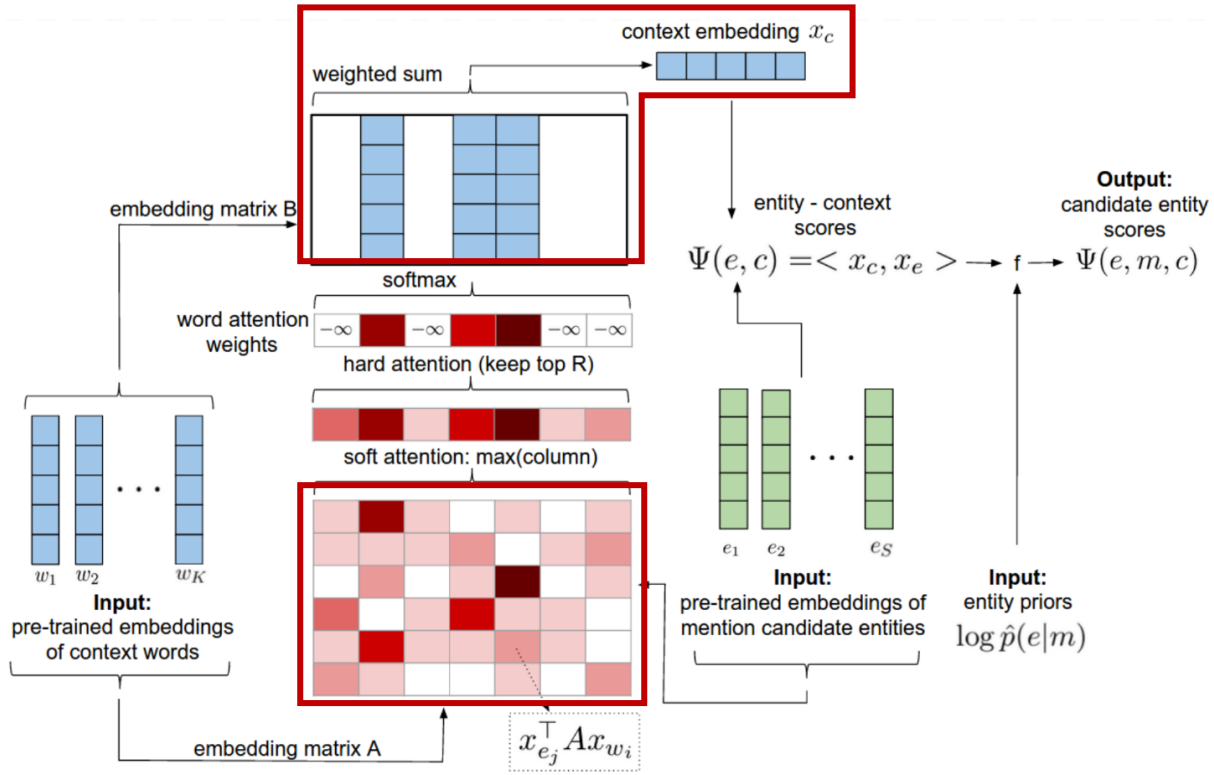
某个提及对应的候选实体集合：

$$\Gamma(m) = \{e_1, \ldots e_s\}$$

- **基于神经网络的实体链接技术**

- 代表性方法：Deep Joint Entity Disambiguation with Local Neural Attention

  - 局部神经注意力计算局部分数



1. 计算上下文中每个词的注意力得分：

$$u(w) = \max_{e \in \Gamma(m)} \mathbf{x}_e^\top \mathbf{A} \mathbf{x}_w$$

2. 计算完后筛选，得到top-R得分的词：

$$\bar{c} = \{w \in c | u(w) \in \text{topR}(\mathbf{u})\}$$

3. 使用softmax，得到注意力权重：

$$\beta(w) = \begin{cases} \frac{\exp[u(w)]}{\sum_{v \in \bar{c}} \exp[u(v)]}, & \text{if } w \in \bar{c} \\ 0 & \text{otherwise.} \end{cases}$$

4. 计算某个候选实体和提及上下文的得分：

$$\Psi(e,c) = \sum_{w \in \bar{c}} \beta(w)\, \mathbf{x}_e^\top \mathbf{B}\, \mathbf{x}_w$$

- **基于神经网络的实体链接技术**

- 代表性方法：Deep Joint Entity Disambiguation with Local Neural Attention

  - 计算文档内所有提及的某候选实体组合的全局分数

$$g(\mathbf{e},\mathbf{m},\mathbf{c}) = \sum_{i=1}^{n} \Psi_i(e_i) + \sum_{i<j} \Phi(e_i, e_j)\ , \quad \Phi(e,e') = \frac{2}{n-1}\,\mathbf{x}_e^{\top}\mathbf{C}\,\mathbf{x}_{e'}$$

组合中实体局部分数之和 **候选实体组合的全局分数** **↑ 基于双线性乘积确保主题一致性**

    - 文档包含的所有提及： $\mathbf{m} = \{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_n\}$
    - 每个提及对应的上下文嵌入： $\mathbf{c} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n\}$
    - 文档可能的候选实体组合 $\mathbf{e}$ 有 $\Gamma(\mathbf{m}_1) \times \Gamma(\mathbf{m}_2) \times \cdots \times \Gamma(\mathbf{m}_n)$ 种可能

- **基于预训练语言模型的实体链接技术**

- 如何提高实体链接的效果?
  - 传统方法：基于Word2vec
    - Static，Bag of words，Token-level
  - 当前主流方法：基于预训练语言模型
    - Context，Bidirectional，Sentence-level

- **基于预训练语言模型的实体链接技术**

- 代表性方法：Scalable Zero-shot Entity Linking with Dense Entity Retrieval (Wu L , et al. EMNLP 2020)
  - 基于Bert微调的两阶段、零样本实体链接方法
  - 零样本实体链接
    - $\varepsilon_{\text{train}}$，$\varepsilon_{test}$ 分别表示训练和测试时的知识库，并且 $\varepsilon_{\text{train}} \cap \varepsilon_{test} = \emptyset$
    - 在训练和测试过程中，文本文档、提及和实体字典的集合是分开的，因此在测试时被链接的实体是训练时没有出现过的

- **基于预训练语言模型的实体链接技术**

- 代表性方法：Scalable Zero-shot Entity Linking with Dense Entity Retrieval
  - 一些预备知识：BERT中常见的分隔符
    - [CLS]：对于文本分类任务，在文本前插入[CLS]符号，其所学到的表征将作为整篇文本语义分类的依据
    - [SEP]：对于输入的两句话，采用[SEP]进行分隔
    - [Ms]/[Me]：提及（Mention）的开始和结束
    - [ENT]：用于标识实体（Entity）

- **基于预训练语言模型的实体链接技术**

- 代表性方法：Scalable Zero-shot Entity Linking with Dense Entity Retrieval
  - 第一阶段：使用 Bi-Encoder 生成候选实体（衡量实体与提及的相似度）
    - 提及编码

$$\tau_m = [\text{CLS}]\text{ctxt}_l\ [\text{M}_s]\ \text{mention}\ [\text{M}_e]\ \text{ctxt}_r\ [\text{SEP}]$$

$$y_m = \text{red}(T_1\ (\tau_m))$$

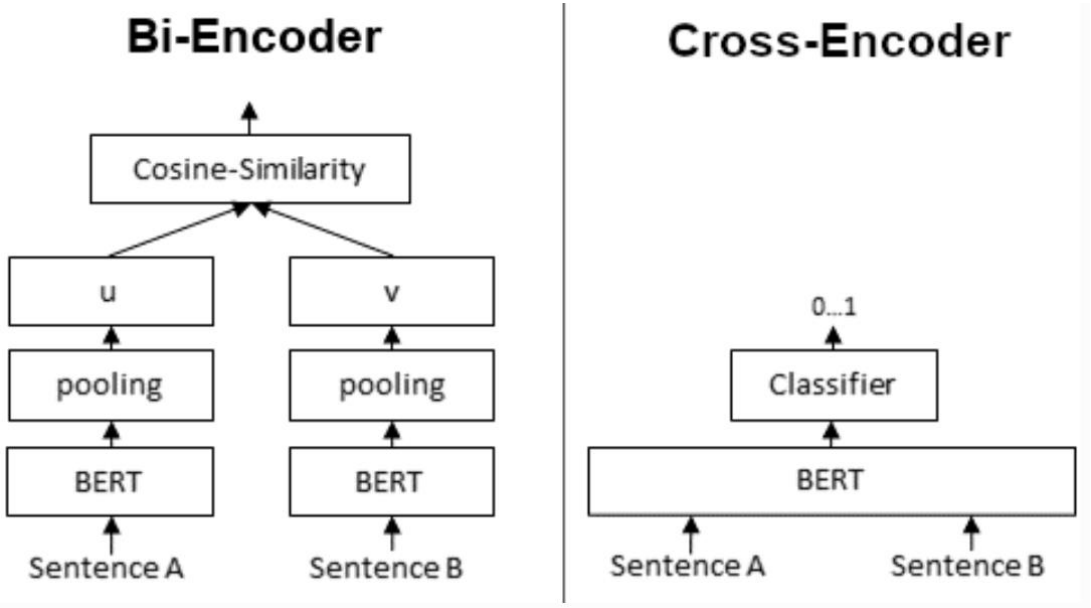
    - 实体编码

$$\tau_e = [\text{CLS}]\ \text{title}\ [\text{ENT}]\ \text{description}\ [\text{SEP}]$$

$$y_e = \text{red}(T_2\ (\tau_e))$$

    - 召回分数（粗筛）

$$s(m, e_i) = \boldsymbol{y_m} \cdot \boldsymbol{y_{e_i}}$$



- **基于预训练语言模型的实体链接技术**

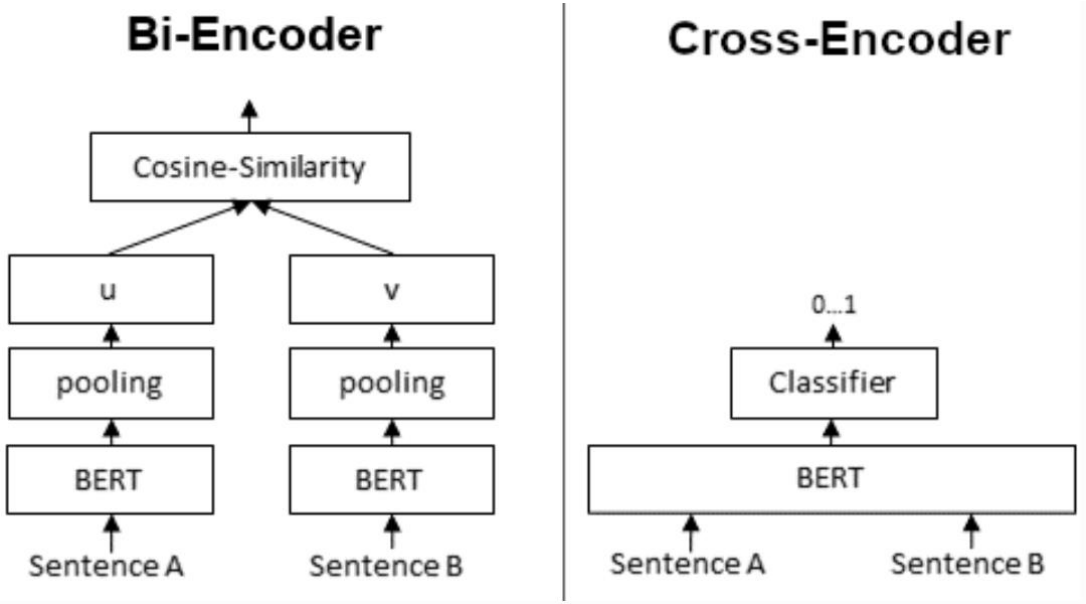
- 代表性方法：Scalable Zero-shot Entity Linking with Dense Entity Retrieval
  - 第二阶段：使用 Cross-Encoder 对候选实体进行排序
    - 提及和实体的整体编码：将提及上下文、实体名称及描述拼接起来

$$\tau_{m,e} = [\mathrm{CLS}]\mathrm{ctxt}_l\ [\mathrm{M}_s]\ \text{mention}\ [\mathrm{M}_e]\ \mathrm{ctxt}_r [\mathrm{SEP}]\text{title}\ [\mathrm{ENT}]\ \text{description}\ [\mathrm{SEP}]$$

$$\boldsymbol{y}_{m,e} = \mathrm{red}(T_{\text{cross}}\ (\tau_{m,e}))$$

    - 精排分数

$$s_{\text{cross}}\ (m,e) = \boldsymbol{y}_{m,e}\boldsymbol{W}$$



- **基于预训练语言模型的实体链接技术**

- 另一种有趣的方案：Efficient One-Pass End-to-End Entity Linking for Questions (BZ Li, et al. EMNLP 2020)
  - 不需要在输入中预先标明提及边界，能够联合进行提及识别和实体消歧
  - 问题定义
    - 给定文档 q 和一组来自 Wikipedia 的实体 $\varepsilon = \{e_i\}$，其中每个实体都有标题 $t(e_i)$ 和对应的文本描述 $\mathrm{d}(e_i)$
    - 目标是输出一个元组 $(e_i, [m_s, m_e])$ 的列表，其中 $e_i \in \varepsilon$ 是实体，对应于由文档 q 中从 m_s 到 m_e 范围 token 组成的提及

- **基于预训练语言模型的实体链接技术**

- 另一种有趣的方案：Efficient One-Pass End-to-End Entity Linking for Questions (BZ Li, et al. EMNLP 2020)
  - 提及识别
    - 实体表征：$\mathbf{x}_e = \text{BERT}_{[\texttt{CLS}]}([\texttt{CLS}]t(e_i)[\texttt{ENT}]d(e_i)[\texttt{SEP}]) \in \mathbb{R}^h$
    - 文档 token 表征：$[\mathbf{q}_1 \cdots \mathbf{q}_n]^\intercal = \text{BERT}([\texttt{CLS}]\, q_1 \cdots q_n[\texttt{SEP}]) \in \mathbb{R}^{n \times h}$
      - 相比于之前少了[Ms]/[Me]的标识，而转为对每个token进行表征

- **基于预训练语言模型的实体链接技术**

- 另一种有趣的方案：Efficient One-Pass End-to-End Entity Linking for Questions (BZ Li, et al. EMNLP 2020)
  - 提及识别
    - 考虑文档中所有长度不超过 L 的子串，为每一个 token 计算三个分数，作为它是提及的开头、结尾或中间的分数（序列标记的思路！）

$$s_{start}(i) = w_{start}^T q_i, \quad s_{end}(j) = w_{end}^T q_j, \quad s_{mention}(t) = w_{mention}^T q_t$$

    - 子串是提及的概率（基于Sigmoid函数）

$$p([i,j]) = \sigma(s_{start}(i) + s_{end}(j) + \sum_{t=i}^{j} s_{mention}(t))$$

- **基于预训练语言模型的实体链接技术**

- 另一种有趣的方案：Efficient One-Pass End-to-End Entity Linking for Questions (BZ Li, et al. EMNLP 2020)
  - 实体消歧
    - 某提及链接到各实体的概率分布

$$\mathbf{y}_{i,j} = \frac{1}{(j-i+1)} \sum_{t=i}^{j} \mathbf{q}_t \in \mathbb{R}^h$$

$$s(e,[i,j]) = \mathbf{x}_e^{\intercal} \mathbf{y}_{i,j}$$

$$p(e|[i,j]) = \frac{\exp(s(e,[i,j]))}{\sum_{e' \in \mathcal{E}} \exp(s(e',[i,j]))}$$

本章小结

知识抽取与表达（上）

- 实体抽取任务
  - 任务定义
  - 基本方法
- 实体对齐
- 实体链接

tongxu@ustc.edu.cn